Haoji Zhang 张颢继

Last Update: May 1, 2026

Homepage GitHub Google Scholar hodge013140@gmail.com Tel (WeChat): +86 18510279299

Research Interests

My research focuses on building efficient and powerful MLLMs, including:

- Efficient Long Video Understanding: AR, VQA, VTG, STVG
- Large-scale MLLM Training: SFT, Post-training (GRPO, Tool-RL)
- Agentic Model and Multimodal Reasoning
- Unified Multimodal Understanding and Generation Model

Education

- **Tsinghua University** *Aug 2024 – June 2027*
  M.S. in *Data Science and Information Technology*, supervised by Prof. Yansong Tang GPA 3.80/4.00
- **Tsinghua University** *Aug 2020 – June 2024*
  B.S. in *Mathematics and Physics* GPA 3.93/4.00 (Top 5%)

Internships

- **JD Retail | TGT Intern** *Jan 2026 – May 2026*
  Worked on unified multimodal models, focusing on alpha-aware transparent image generation and editing.
- **ByteDance Intelligent Creation | Jindouyun Intern** *Feb 2025 – Jan 2026*
  Worked with Dr. Longyin Wen on agentic MLLM, long video understanding and spatio-temporal grounding.
- **ByteDance Seed Research | Research Intern** *Sept 2023 – Dec 2024*
  Worked with Dr. Jiashi Feng and Dr. Xiaojie Jin on efficient MLLM and long video understanding.

Selected Publications & Preprints

- **Thinking With Videos: Multimodal Tool-Augmented RL for Long Video Reasoning** **CVPR 2026**
  Haoji Zhang, *Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Bowen Zhang, Zhichao Zhou, Dongliang He, Yansong Tang*
  We propose the first agentic video reasoning RL framework, achieving new state-of-the-art on many long video VQA and temporal grounding benchmarks with multimodal Chain-of-thought.
- **Thinking With Bounding Boxes: Enhancing Spatio-Temporal Video Grounding via RFT**
  Xin Gu,* ***Haoji Zhang****, Qihang Fan, Zhipeng Zhang, Libo Zhang, Guang Chen, Fan Chen, Longyin Wen, Sijie Zhu*
  We developed STVG-o1, a state-of-the-art spatio-temporal video grounding MLLM using "bounding-box chain-of-thought" and GRPO post-training.
- **Vidi2: Large Multimodal Models for Video Understanding and Creation** **Tech Report**
  Haoji Zhang*, Chia-Wen Kuo*, Dawei Du*, Sijie Zhu*, Xin Gu*, Zhenfang Chen*, Longyin Wen, Xiaohui Shen, et al.*
  Vidi2 is a strong MLLM base model with fine-grained spatio-temporal grounding and temporal retrieval ability for long video understanding and editing.
- **Flash-VStream: Memory-Based Real-Time Understanding for Long Video Streams** **ICCV 2025**
  Haoji Zhang, *Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, Xiaojie Jin*
  We propose Flash-VStream, a video MLLM for real-time long video understanding, outperforming Qwen2-VL on multiple long video QA benchmarks. Our model won the Champion of CVPR 2024 LOVEU Challenge.
- **Uni-AdaFocus: Spatial-Temporal Dynamic Computation for Video Recognition** **TPAMI 2025**
  Yulin Wang,* ***Haoji Zhang****, Yang Yue, Shiji Song, Chao Deng, Junlan Feng, Gao Huang*

We propose Uni-AdaFocus, a spatial/temporal/sample-wise dynamic network for efficient video action recognition, achieving a 4.8× speedup with comparable accuracy.

- **Ponder & Press: Advancing Visual GUI Agent towards General Computer Control** **ACL 2025**
  Yiqin Wang, **Haoji Zhang***, Jingqi Tian, Yansong Tang*
  We propose Ponder & Press, a divide-and-conquer GUI agent framework that only relies on visual input to mimic human-like interaction with GUI OS, including Android, PC and web browsers.
- **Self-Calibrated CLIP for Training-Free Open-Vocabulary Segmentation** **TIP 2025**
  *Sule Bai, Yong Liu, Yifei Han, **Haoji Zhang**, Yansong Tang*
  We propose SC-CLIP, a training-free open-vocabulary segmentation framework that achieves competitive performance on various segmentation tasks.
- **PREIM3D: 3D Consistent Precise Image Attribute Editing from a Single Image** **CVPR 2023**
  *Jianhui Li, Jianmin Li, **Haoji Zhang**, Shilong Liu, Zhengyi Wang, Zihao Xiao, Kaiwen Zheng, Jun Zhu*
  We propose PREIM3D, a novel framework for 3D-aware image attribute editing that achieves better 3D consistency and precision at large camera poses.

Honors & Awards

- First-Class Outstanding Scholarship of Tsinghua University (Top-10%) *2021 / 2022 / 2023 / 2025*
- Outstanding Bachelor Graduate of Beijing (Top-5%) *2024*
- **Champion** (among 20+ teams) of LOVEU@CVPR'24: Long-term VQA Challenge *2024*
- **Honorable Mention**, Mathematical Contest in Modeling (MCM) *2022*
- **Gold Medal** (20^{th} in China), THUWC Tsinghua University Winter Camp in Informatics *2019*
- **Silver Medal** (59^{th} in China), NOIWC National Olympiad in Informatics Winter Camp *2019*
- **Bronze Medal** (154^{th} in China), NOI National Olympiad in Informatics *2019*

Academic Service

- Conference Reviewer for ICCV, AAAI, ICLR, CVPR, NeurIPS
- Journal Reviewer for JVCIR

Skills

- Programming Languages: Python, C, C++, Rust
- Deep Learning: PyTorch, Transformers, Diffusers, DeepSpeed, Ulysses, TRL, VeRL, vLLM
- Web Technologies: HTML/CSS, JavaScript, Vue, Flask, Nginx
- Languages: Chinese (native), English (fluent, TOEFL: 108 / 120, GRE: 329 / 340)